# Skin Cancer Image Classification

36320 , 29921, 35190

January 20, 2022

## Abstract

With the onset of the third of era of globalisation in 1989, the world has gone through massive changes with the help of technology. In order to promote this advancement and to make a significant contribution to the world, especially to the health sector, we have designed models to classify skin cancer images into Malignant and Benign subgroups and then delve deeper into specific types of cancer. First, a general model that classifies the cancer images into two large groups stated above, is constructed. Next, this general model is applied to new baseline data to determine its suitability and robustness using different metrics. A similar procedure is replicated for examining specific cancer types. The drawbacks and possible improvements of the models have also been highlighted. For the purpose of our study we have applied Convolutional neural network (CNN) on two real world data-sets using Artificial Intelligence. .

**Keywords:** Artificial Intelligence, Neural Networks, Convolutional Neural Network (CNN), Image Classification

## 1. Introduction

In light of the present Covid-19 pandemic where pharmaceutical companies and health care systems globally have left no stones unturned to reduce the levels of infection around the world by developing vaccines at a fast pace, applying such a model in assessing patients' stage of a disease will aid the medical system. The primary reason for the same is that there have been instances where patients with serious diseases like cancer have been neglected because of the overburdening of healthcare facilities, frequent lockdown's and to avoid such critical patients from catching the virus, thereby leading to more damage than cure.

Just like the numerous and never-ending strains of Coronavirus, skin cancer too appears when varied mutations occur in the DNA of skin cells (Craythorne and Al-Niami, 2017). These mutations cause the healthy cells to grow out of control and form a mass of cancer cells, which results in a tumour if not detected at the right time. (Chalap and Al-Awsi, 2019; Oberholzer et al., 2012). This tumour that occurs can be malignant (skin cancer assaults neighboring tissues like lymph nodes or when invading other organs via the bloodstream, resulting in metastasis) or benign (unable to invade neighbouring tissues and can be removed by surgery) (Chalap and Al-Awsi, 2019). Thus, early detection of skin cancer is critical (Bickers et al., 2006), given that at present it is the most common cancer (Craythorne and Al-Niami, 2017; Guy Jr et al., 2015) and also in the event that it is declared malignant, it becomes even more imperative that an effective and imminent diagnosis is undertaken in order to give the patients the best chance of successful skin cancer treatments by virtue of radiation and chemo-therapies. (Koh, 1991), This will further not only improve their quality of life but also help in reducing the death rates by such diseases. (Diepgen and Mahler, 2002; Trager et al., 2020). For this early identification, experts use images of the patient's skin to diagnose and compare them with the different pictures of skin cancers, which are located in different skin cancer image data sets (Chuchu et al., 2018; Tschandl et al., 2020). The great variety of skin lacerations and their configuration in images makes their classification a difficult task (Dubal et al., 2017)

To reduce this tedious process of detection and succour the medical system we suggest that the department use automated classification using artificial intelligence, with the help of neural network modelling, thereby assisting the medical staff in effectively classifying images of different types of cancer. The first approach to the problem is to develop a model that simply classifies skin cancers into benign and malignant. The lesions classified as malignant are basal cell carcinoma (bcc) and melanoma (mel). The rest of the lesions are classified as benign: actinic keratoses and intraepithelial carcinoma /Bowen's disease (akiec), benign keratosis-like lesions (solar lentigines /seborrheic keratoses and lichen-planus like keratoses (bkl), dermatofibroma (df), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage (vasc). A more specific model is then established to classify the different skin cancers by type which the model will identify by their abbreviations (mel, bcc, akiec, bkl, df, nv, vasc). Although there are many approaches to this classification problem (Dubal et al., 2017; Esteva et al., 2017; Goyal et al., 2020; Lau and Al-Jumaily, 2009; Maron et al., 2019), our contribution is that after training the models and assessing them on the same data set, the performance of the models are tested on new data sets, in which image form may vary, and the results are then evaluated. This in turn enables us to check the robustness of the respective models, and thus ultimately determine whether they could be of real use to medical-experts.

## 2. Data

In this study three data bases are deployed. The reference data base can be found in `https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000`. Our two models to classify the cancer images will be constructed from this data base. The data set includes 10015 images with the labels of the different types of skin cancer lesions, as stated in the Introduction.

When applying the generic model, our interest is to classify the lesions as malignant and benign, hence data pre-processing needed to be done, in order to set the labels of benign and malignant, depending on the type of skin lesion. The lesions classified as malignant are basal cell carcinoma (bcc) and melanoma (mel) and the rest were classified as benign, as stated in the Section 1. One of the problems that arose when setting the column with the new malignant or benign labels, is that most of the lesions are benign, hence leading to a large imbalance between the two class, making the investigation cumbersome but nevertheless challenging enough to achieve our goals promptly. For the specific model the classes are also imbalanced, since the class counts ranges from 115 (class: 'df') to over 6700 (class: 'nv'). The different class for the general and specific models respectively, can be visualized in Figure 3 and Figure 4. Before fitting and applying the models, the data was divided into training, validation and testing, with the following ratios: 70% training and 30% testing, where 30% of the training data was used for validation.

After applying both the models to the reference data set, the general model will be applied to the data found in `https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign/` and the specific model to the data `https://www.kaggle.com/nodoubttome/skin-cancer9-classesisic`. These two data sets mentioned above contains the labels of benign and malignant, and the type of skin lesions, respectively. However, in this case all the images are used for testing purposes. Furthermore, in the new data of the specific model, the squamous cell carcinoma class of images were eliminated as this class did not appear in the training model and the class of pigmented benign keratosis and seborrheic keratosis were merged together as they appear as one class in the training model (bkl).

## 3. Methods

Before being able to apply the methods to the real data there are some methods and hyper-parameters that needs to be discussed. A useful technique for data pre-processing is Data Augmentation, which consists of the process of generating new synthetic samples for training purposes from the available images by applying tools such as: zoom, rotation, horizontal and vertical translation, shearing, adjusting image brightness, colour shift and horizontal and vertical flips. (Perez and Wang, 2017). It is pertinent to note here that since there exists a skewed distribution of classes, it was essential to assign different weights to the classes to reduce the discrepancyin our models.(Zhu et al., 2018).

The Artificial Intelligence algorithm that will be used is the Convolutional neural network (CNN), which is a type of neural network specializing in image classification (Albawi et al., 2017). It consists of a sequence of multiple hidden layers which normally are convoluted layers followed by activation layers (Albawi et al., 2017; Guo et al., 2017).

We further discuss about the hyper-parameters that we have taken into account for the scope of our study. The learning rate is one such criteria that we used as it limits the rate at which the algorithm learns or updates the values of the variables in under observation. (Fang et al., 2005). The second one is the dropout rate which is applied to the layers of the Neural Network and it randomly zeroes out a fraction of the neuron values during the training of the model. This prevents multiple layers from extracting the similar features from the data, and hence Helps avoid overfitting (Srivastava, 2013). The third, is the batch size is that specifies the number of observations that the algorithm will go through before updating the characteristics of the model (Brownlee, 2018), Finally the number of epoch is a framework that determines how many times the learning algorithm will go through the entire data set before updating the parameters (Brownlee, 2018). This epoch consists of one or more batches (Brownlee, 2018). The number of nodes or neurons in each layer can also be specified, this category is also known as the width of the neural network (Karsoliya, 2012).

All of the aforementioned can be tuned in order to find their ideal values, which differ dependent on the data and model formulation. In our case to tune both General and Specific models, we use *Hyperband* tuning, a novel tuning method proposed by Li (2018). Due to the large computational power and time required to perform *Grid Search*, as well due to the wastefulness of *Random Search* in which the tuner may fully train models with parameter choices which can be seen to be a bad picks from the start, neither of the two tuning methods were ideal choices. This is where *Hyperband* comes into play, as a solution to the aforementioned problems. It operates by randomly sampling all the combinations of hyperparameters, and using these configurations, it trains the model for a few epochs (less than the maximum number of epochs i.e *max_epochs* parameter), and chooses the best candidates. This is done iteratively (one can control the number of iterations done), and full training and evaluation using the maximum amount of epochs is used on the final chosen candidates.

In supervised learning algorithms there is always a loss function and a method to minimize this function during the training process (Salimans et al., 2016). In our case the loss function used in the general model is the *Binary Cross Entropy* as it is a binary classification task and the activation function employed in this case is the *Sigmoid Activation Function* as it is the only one compatible with this loss function and can be used to transform continuous space values to binary ones (Creswell et al., 2017). In the case of the specific model the loss function used is *Categorical Cross Entropy* aswe have a multi-class classification task and the activation function is *Softmax Activation Function* which is the recommended function for in this case. It transforms a vector of numbers into a vector of probabilities, where these probabilities are proportional to the relative value of the entries of the vector (West and O'Shea, 2017). The optimizer used for these loss functions is the *Adam Optimizer* as it intuitively combines the *Gradient Descent with Momentum Algorithm* and the *RMSP Algorithm* and is a very efficient method when working with a plethora of parameters or large data-sets. (Klooster, 2021). Further, it is able to limit the rate of gradient descent such that it is able to ”pass” the local minima while having minimum oscillation when a global minimum is reached (Elderman, 2019). In this way it is able to outperform other optimizers like *AdaGrad*, *RMS Prop*, *SGD Nesterov* and *AdaDelta* (APPENDIX) (Klooster, 2021; Metz et al., 2018). Additionally, in the specific model, the binary *Accuracy* metric, formulated in Section 4.1.1, is used in the validation process to choose the optimal threshold for the posterior testing.

# 4. Metrics

## 4.1 General model

### 4.1.1 Confusion Matrix

The confusion matrix summarizes the performance of the algorithm in a binary classification problem (Hossin and Sulaiman, 2015; Kohl, 2012; Raschka, 2014). Let $P$ be the label of the *Positive* class and $N$ the label of the *Negative* class (Canbek et al., 2017; Powers, 2020; Raschka, 2014). The *Confusion matrix* is represented by the following table (Kohl, 2012; Luque et al., 2019; Parikh et al., 2008; Powers, 2020; Tharwat, 2020):

| | | Actual | | |
|---|---|---|---|---|
| | | Positive (Malignant) | Negative (Benign) | Total |
| Predicted | Positive (Malignant) | $TP$ | $FP$ | $TP + FP$ |
| | Negative (Benign) | $FN$ | $TN$ | $FN + TN$ |
| | | $P = TP + FN$ | $N = FP + TN$ | |

The rows of the table represent the predicted outcomes and the columns the actual class (Hossin and Sulaiman, 2015; Powers, 2020; Raschka, 2014). It has four possible outcomes, the first diagonal represents the samples that were correctly classified as *Positives* and *Negatives*, conformed by the *True Positives (TP)* and the *True Negatives (TN)*, respectively (Canbek et al., 2017; Parikh et al., 2008; Tharwat, 2020). The other diagonal represents the observations that were misclassified as *Positives* and *Negatives* conformed by the *False Positives (FP)* or *Type I error* and the *False Negatives (FN)* or *Type II error*, respectively (Parikh et al., 2008; Powers, 2020). In our case, malignant lesions will be represented by the *Positives* and the benign lesions will be represented by the *Negatives. Type II errors* can be worse than *Type I errors* in skin cancer detection, as they are malignant cancers that are not detected. From the *Confusion matrix* many other classification metrics can be computed (Canbek et al., 2017; Hossin and Sulaiman, 2015; Luque et al., 2019):

$$Sensitivity = \frac{TP}{TP + FN} \qquad FPR = 1 - Specificity = \frac{FP}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad G{-}\text{Mean} = \sqrt{Sensitivity \cdot Specificity}$$

$$Precision = \frac{TP}{TP + FP} \qquad F_1{-}\text{Score} = \frac{2 \cdot Precision \cdot Sensitivity}{Precision + Sensitivity}$$

Every metric lies between $[0, 1]$ (Canbek et al., 2017; Luque et al., 2019), where 1 leads to perfect prediction and 0 corresponds to worst possible prediction (Chicco and Jurman, 2020; Tharwat, 2020),except in the case of the $FPR$ which is the opposite (Tharwat, 2020). The *Accuracy, Precision* and $F_1$*-Score* are sensitive to imbalanced data sets and can give misleading results as they provide an optimistic measure for the capability of the method in detecting the majority class (Luque et al., 2019; Tharwat, 2020). While the *Sensitivity*, the *Specificity*$(1 - FPR)$ and the *G-Mean*, which is the combination of the two, are robust to imbalanced data (Hossin and Sulaiman, 2015; Luque et al., 2019; Raschka, 2014; Tharwat, 2020).

### 4.1.2 ROC curve

The *ROC* or Receiver operating characteristic curve is a plot of the True Positive Rate ($TPR$ or Sensitivity) against the False Positive Rate ($FPR$) for different thresholds of a parameter (Hoo et al., 2017; Marzban, 2004). The area under this curve, also denoted as *AUC* can be an additional metric to test the models (Hoo et al., 2017).

Hence to evaluate the models the *AUC* can be used. A model with *AUC* $= 0.5$ represented by a diagonal line in the *ROC* plot, corresponds to a model which classifies at random (Hoo et al., 2017; Marzban, 2004). A model with *AUC* $= 1$, corresponds to a model which perfectly predicts the classes of different images (Hoo et al., 2017; Marzban, 2004). As a result, the approach which yields the highest *AUC*, corresponding to the model with an *ROC* curve closest to the top left corner, is preferred (Hoo et al., 2017).

## 4.2 Specific model

### 4.2.1 Confusion Matrix

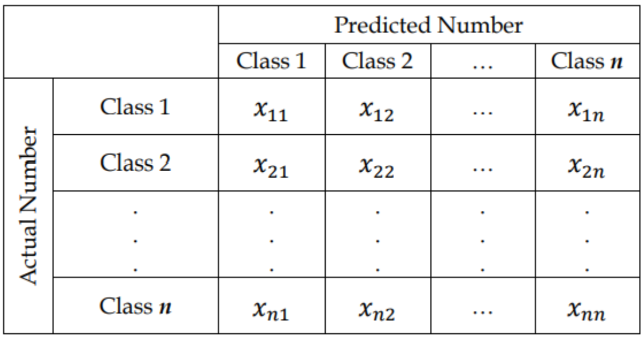
The confusion matrix for the multi-class problem will be the one represented in Figure 1.1(Deng et al., 2016; Manliguez, 2016). The number of classes $n$ will be seven in our case and $x_{ij}$ represents the number of images classified as $Class_j$ but belonging to $Class_i$ (Deng et al., 2016). Now the error for each class will be $\sum_{i \neq j} x_{ij}$.

From the *Confusion matrix* many other classification metrics can be computed for class $i$ using the one against all approach (Deng et al., 2016). These metrics have similar properties to the ones for the binary case as by using this one against all approach, we are transforming the metrics for class $i$ into a binary classification problem, where the all the classes $j \neq i$ are merged into one class. In this case, the *Accuracy* for class $i$ will be the same as the *Sensitivity* for class $i$, hence we omitted it. The metrics for class $i$ are defined as follows (Deng et al., 2016):

$$Sensitivity(i) = \frac{x_{ii}}{\sum_{j=1}^{7} x_{ij}} \qquad FPR(i) = 1 - Specificity(i) = \frac{\sum_{i \neq j} x_{ji}}{\sum_{i \neq j} \sum_{k=1}^{7} x_{jk}}$$

$$G\text{−Mean}(i) = \sqrt{Sensitivity(i) \cdot Specificity(i)}$$

$$Precision(i) = \frac{x_{ii}}{\sum_{j=1}^{7} x_{ji}} \qquad F_1\text{−Score}(i) = \frac{2 \cdot Precision(i) \cdot Sensitivity(i)}{Precision(i) + Sensitivity(i)}$$



| | | Predicted Number | | | |
|---|---|---|---|---|---|
| | | Class 1 | Class 2 | … | Class *n* |
| Actual Number | Class 1 | $x_{11}$ | $x_{12}$ | … | $x_{1n}$ |
| | Class 2 | $x_{21}$ | $x_{22}$ | … | $x_{2n}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | Class *n* | $x_{n1}$ | $x_{n2}$ | … | $x_{nn}$ |

Figure 1: Confusion Matrix for multiple classes

### 4.2.2 Macro Averages

In order to get an overall result of the metrics for each class, we compute the *Macro-Average* of each metric by computing the mean of the specific metric. In this way we can obtain *Macro-Sensitivity, Macro-FPR, Macro-G-Mean, Macro-Precision* and *Macro-$F_1$-Score* (Grandini et al., 2020). Finally the *Accuracy* of the model can be defined as follows (Deng et al., 2016):

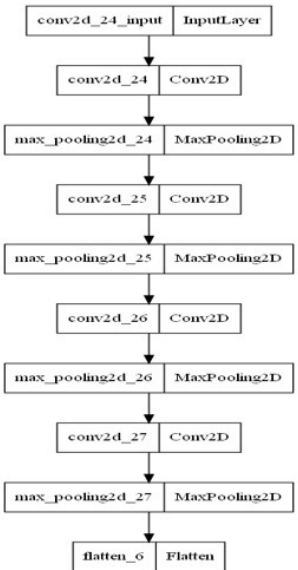
$$Accuracy = \frac{\sum_{i=1}^{7} x_{ii}}{\sum_{i=1}^{7} \sum_{j=1}^{7} x_{ij}}$$

### 4.2.3 ROC Curve

The *ROC* Curve for the multi-class problem will be very similar to the *ROC* for the binary case, but now we will have multiple *ROC* Curves, one for each class. Then, we will plot for class $i$, the *Sensitivity(i)* against the *FPR(i)* for different thresholds of a parameter (Marzban, 2004). Hence, to evaluate the model, we can use the area under the $ROC(i)$ Curve ($AUC(i)$) for class $i$, and compute the *Macro-AUC* by taking the average *AUC* of all the classes (Deng et al., 2016).

## 5. General Model

As discussed in the Introduction, we are building a generic model to classify the lesions as 'malignant' and 'benign'. We initially built a basic model, the results of which were inconclusive, thereby resulting in inaccurate predictions where all test images were classified into one class. The reasons for this occurrence is the large imbalance in our data-sets already stated in Section 2, as well as Keras' (Python interface used for building artificial neural networks) automatic assumption that the ideal threshold for classification is 0.5. This model was thus disregarded, but can be considered an important stepping stone in realising the necessity of finding more ideal parameters and in particular the right classification threshold.



Thus, a tuned model is built on the existing data-set with architecture as displayed in Figure 2. The parameters such as learning and dropout rates are adjusted; and then we test this model on a new data-set to assess the results. The purpose of this model is to help the healthcare system globally to be able to analyze whether a particular disease is cancerous or not early on and thereby provide the best possible remedy to the concerned patient, consequently increasing the life expectancy of individuals around the world.

Additionally, as mentioned previously an issue we had to tackle with was imbalanced classification, given 8902 images are classified as Benign and the rest 1113 images as Malignant (Figure 3). To tackle this imbalance and to be able to make a steady model, we apply data augmentation to the training images as described in the first paragraph of Section 3. The augmented images are then added to the training folders of the malignant and benign diseases, increasing the total number of training images by approximately 260%. All testing, training and validation images are further re-scaled with a factor of 1./255, transforming every pixel value from range [0,255] to [0,1], in order to have equal treatment of images of high and low pixel range, as well as to enable the use of a typical learning rate.

Moreover to reduce, the tendency of simply classifying the test images into the majority class 'benign' when performing model predictions, we calculate class weights and use them in model fitting. The class weights obtained are as follows:

$$\text{benign} : 0.59, \ \text{malignant} : 3.07$$

Here the class with the highest frequency i.e Benign is assigned the lowest weight.

### 5.1 Model Building

A Convolutional Neural Network is built to make our predictions. Further, to reduce the effect of such a large imbalance in the data, we try and identify the optimal hyper-parameters using *Hyperband* tuning as described in Section 3. When developing the tuned model, one must define the hyper-parameters to be tuned, which in this case were chosen to be the learning rate, the dropout rate, the number of units in the second to last dense layer and the classification threshold with possible values [0.01,0.001,0.0001], [0.0,0.1,0.2] [32,64,128] and a range of 15 values between [0.53,0.55], respectively. The *Binary Crossentropy* loss function and *Adam Optimizer* were employed.

Using the maximisation of the Validation *Binary Accuracy* as objective, the tuner identified the optimal hyper parameters to be Learning rate: 0.0001, Dropout Rate: 0, Number of Units: 64 and Threshold value: 0.54714

Using these parameters the model was trained for 20 epochs and evaluated on the testing data. It is important to note here that following predictions, we use the obtained threshold value 0.54714, in order to achieve an appropriate test image predicted class index.

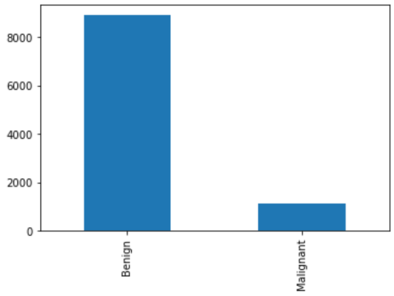


Figure 3: Counts of malignant v benign skin lesions in base/reference data set

### 5.2 Results

Below in Figure 4, one can see the ROC Curve that was generated while making predictions.

The Receiver Operating Characteristic curve (or ROC curve) has an AUC value of 0.81. The AUC value lies between 0.8 to 0.9, significantly superior to 0.5. This might suggest that our model is accurately able to diagnose whether a person is suffering from a Malignant or Benign skin lesion on the basis of an image.Moreover, the accuracy value of our model is 83.34 percent and the binary accuracy is 72 percent, which is another good indication of an well developed model.

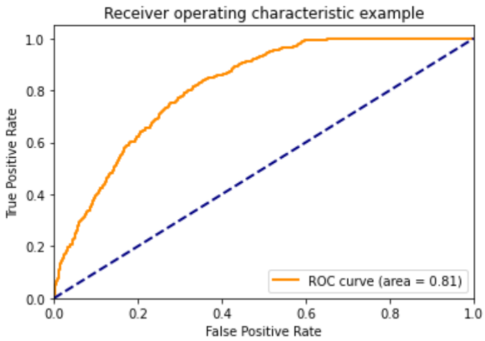


Figure 4: ROC Curve for General Model

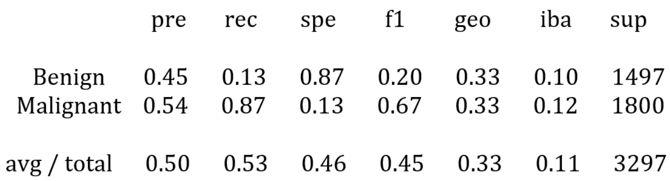
Moving on to the classification report, the macro average precision score of both the classes is 0.84 and the geometric mean is 0.73, which is considered to be excellent from an overall standpoint, but as one drills down to each of the classes, it is evident that the respective values in the case of Benign are higher than Malignant by a very high margin *Appendix-A Classification Report.* However at the same time we observe that rest of the parameters are all well balanced for both the classes, as a result indicating an overall well-performing model.

## 6. Application of General model to unseen data set

Once the General model is constructed let us apply the model to the new data set containing the labels for 'benign' and 'malignant' type of cancers described in Section 2. The models are applied now to new data, which which may be different to the images used for the training and so this can result in a harder task for the model to adequately predict the true label of the images.

### 6.1 Initial Model

The results of the model after trying to predict for a new data set are displayed in Figure 5 and Figure 6. It can be seen there is a clear drop in most of the metrics when testing on this new data set, having approximately an average of *Precision*, *Recall*, *Specificity*, $F_1$*-Score* and *G-Mean* a decrease of 40%, 27%, 37%, 41% and 55%, respectively. Although the *Recall* and the *Precision* decrease, they have increased for the malignant class quite significantly, implying that the performance has not only improved in detecting malignant cancer images but also its capacity in reducing the number of misclassified malignant images. This is the reason behind the high $F_1$*-Score* for the malignant class



```
                pre    rec    spe    f1     geo    iba    sup

    Benign      0.45   0.13   0.87   0.20   0.33   0.10   1497
 Malignant      0.54   0.87   0.13   0.67   0.33   0.12   1800

avg / total     0.50   0.53   0.46   0.45   0.33   0.11   3297
```

Figure 5: General Model applied to new data set Classification report

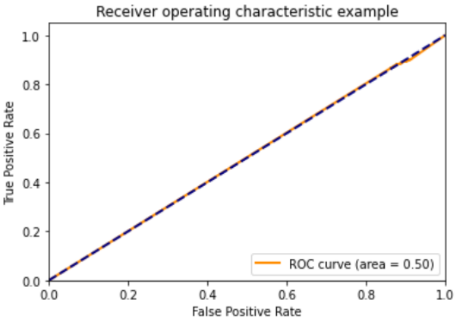


Figure 6: ROC Curve for General Model applied to new data set

and the high *Specificity* for the benign class, as it is equivalent to the *Sensitivity* of the malignant class. Nevertheless, the *Recall* from the benign class has dropped from 0.73 to 0.13 which is worse than employing a random classifier and makes the value of the *G-Mean* very low. The *Accuracy* in this case is 0.5323 and can be taken into account as it is a balanced data set. Finally, the *AUC* is 0.5 which implies that we will get the same *Accuracy* as random classification, this is far from ideal.

### 6.2 Problems and tricks to mitigate the performance drop

Predicting the labels for images of new data is a hard problem as the angle, the size, the lighting, the skin color and the affected area can be different, despite our efforts of tackling this using data augmentation. It is especially challenging for images of skin lesions that are lighter, that have not properly developed yet(PICTURE IN THE APPENDIX). Even though they are from the same background the images used for training and the images from the new data may not have the same features nor the same feature distribution. One way to mitigate this problem is to eliminate possible outliers and through feature selection to avoid the model focusing on irrelevant details, which leads to overfitting.

One of the reasons that can explain having low *Sensitivity* of the benign class (or the *Specificity* of our model) can be due to the fact that the new data set is balanced, however in the trained model we gave more weight to the malignant class (minority class in training model) as the model was trained to reduce the misclassification error of the minority class (malignant). By applying it to the new data set will still try to reduce this error but in this case there are many more malignant pictures. A possible solution will be to train the model on this balanced data set or training the model on the imbalanced data set using the methods of oversampling (increases the number of samples of the minority class, can lead to overfitting), undersampling (decreases the number of samples of the majority class, can lead to loss of information) or ensemble learning techniques like Bagging and Boosting, which lead to more stable results.

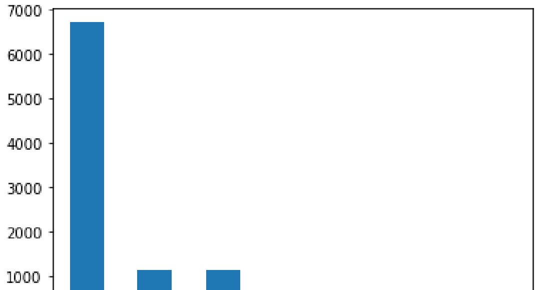
The model might not be trained enough for the classes and this can be mitigated by generating synthetic samples using techniques such as Data Augmentation which generates additional training data. It was however used when training the models, and although it improved the performance for the testing data, the performance decreased significantly when applying it to a new data set. This technique together with the Dropout rate (applied to a layer of the Neural Network it randomly drops output units during the training of the model), prevent the model from learning irrelevant details or noises from the training images, resulting in overfitting. Increasing this Droptout rate in the training phase may lead to better performance when applying the model to a new data set. This overfitting can negatively impact the model when predicting labels for unseen images.

The complexity of the model can also lead to overfitting and to having a bad performance when applied to new data. To avoid overfitting, the model should be tuned but such that it remains broad enough to be able to classify correctly unseen images from new data. In our case, the complexity of CNN is more suitable for the data used in the training and testing rather than for the new data (given the similar features), which results in bad performance for the model. Hence, there might be a problem when selecting the hyperparameters of the model, being suitable for the training data but not for this new data. A simpler choice of the model when selecting and tuning the hyperparameters might be better. This simpler model can be accomplished by reducing the number of epochs (line plots could help in choosing this parameter), batch size, depth of the neural network and the number of nodes per layer. Also increasing the learning rate can mitigate this problem, as low learning rates lead to fit the noise instead of the data structure.

As we saw above the Adam optimizer outperform other optimizers in terms of training cost, however it might be worth trying out other optimizers to see if the performance when testing the model on a new data set is improved.

## 7. Specific Model

Having developed a general model capable of classifying lesions into 'malignant' or 'benign', as well as evaluated its performance on a completely new and unseen data set, we now delve even deeper and try and create a more detailed model, namely the Specific Model. This will have the capability of predicting the specific types of skin lesions found in the images, not only their stage of 'malignant' or 'benign'. Such a model would be of



great use to doctors world wide, as having the knowledge of the specific type of malignant or benign skin disease a person is suffering from, allows for much more effective treatment, and less need for further tests to determine the nature of each given lesion.

As mentioned in Section 1, the base data set contains images of seven classes, namely 'bcc', 'mel', 'akiec', 'bkl', 'df', 'nv' and 'vasc' (details of diseases they represent can be seen in the Section 1, all of varying counts as seen in Figure 5.

To once again counter this imbalance as well as help make the model to be fit, more generalisable, following the split of the data set into training, testing and validation images, data augmentation of the training images is employed. The same methods were used as in the General Model, as described in Section 3, and had the effect of increasing the total number training images from 4907 to 17665 images, a significant change. All testing, training and validation images were then rescaled with a factor of 1./255. As an extra measure of tackling imbalance, weights are again assigned, in order to achieve a much more equal treatment of classes. The class weights are the following:

$$\text{akiec} : 4.34,\ \text{bcc} : 2.79,\ \text{bkl} : 1.29,\ \text{df} : 12.19,\ \text{mel} : 1.28,\ \text{nv} : 0.21,\ \text{vasc} : 10.47$$

where the classes with the highest frequency are assigned the lowest weights.

## 7.1 Initial Model

### 7.1.1 Model Building

Having explored with basic and more advanced model configurations when building the Convolutional Neural Network, an initial model is developed with the architecture displayed in Figure 2, with a Learning rate of 0.001, Adam Optimizer, *Categorical Cross Entropy* Loss function, 50 epochs and batch size of 128. A Dropout rate of 0.2, as well as early stopping monitoring Validation *AUC*, are employed to prevent overfitting and stop the training process when the performance of the model on the validation set starts to deteriorate.

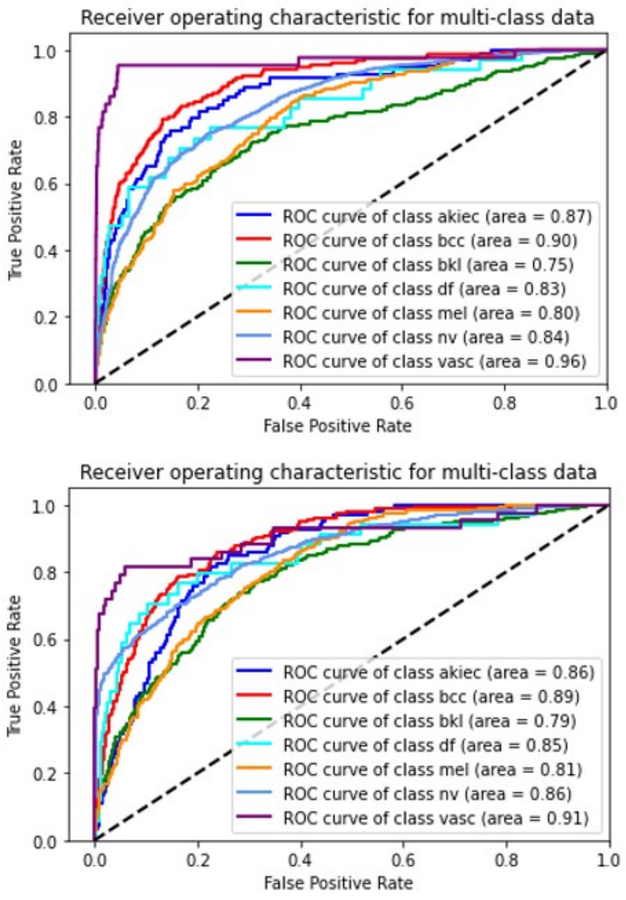
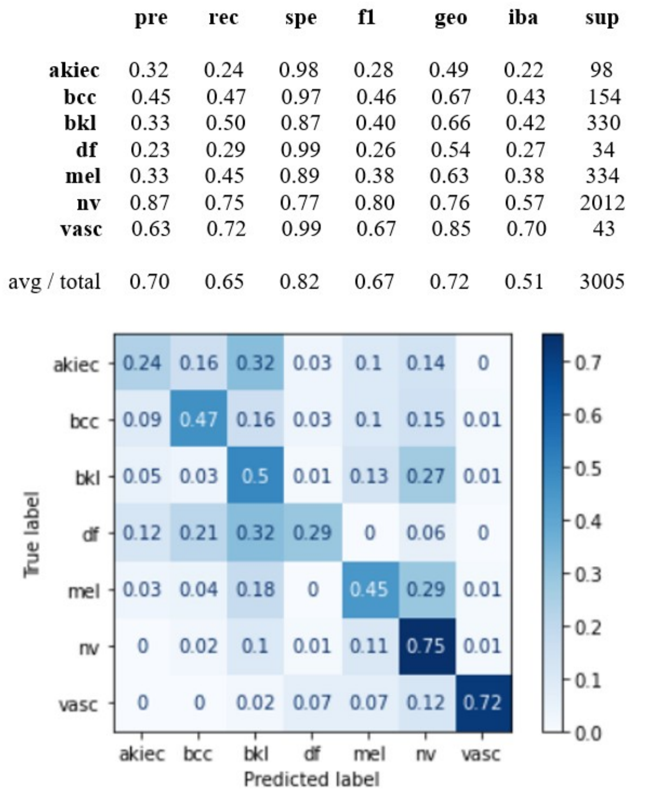


Figure 8: ROC of initial (top) and tuned (bottom) specific models



| | pre | rec | spe | f1 | geo | iba | sup |
|---|---|---|---|---|---|---|---|
| **akiec** | 0.32 | 0.24 | 0.98 | 0.28 | 0.49 | 0.22 | 98 |
| **bcc** | 0.45 | 0.47 | 0.97 | 0.46 | 0.67 | 0.43 | 154 |
| **bkl** | 0.33 | 0.50 | 0.87 | 0.40 | 0.66 | 0.42 | 330 |
| **df** | 0.23 | 0.29 | 0.99 | 0.26 | 0.54 | 0.27 | 34 |
| **mel** | 0.33 | 0.45 | 0.89 | 0.38 | 0.63 | 0.38 | 334 |
| **nv** | 0.87 | 0.75 | 0.77 | 0.80 | 0.76 | 0.57 | 2012 |
| **vasc** | 0.63 | 0.72 | 0.99 | 0.67 | 0.85 | 0.70 | 43 |
| avg / total | 0.70 | 0.65 | 0.82 | 0.67 | 0.72 | 0.51 | 3005 |

| True label \ Predicted label | akiec | bcc | bkl | df | mel | nv | vasc |
|---|---|---|---|---|---|---|---|
| akiec | 0.24 | 0.16 | 0.32 | 0.03 | 0.1 | 0.14 | 0 |
| bcc | 0.09 | 0.47 | 0.16 | 0.03 | 0.1 | 0.15 | 0.01 |
| bkl | 0.05 | 0.03 | 0.5 | 0.01 | 0.13 | 0.27 | 0.01 |
| df | 0.12 | 0.21 | 0.32 | 0.29 | 0 | 0.06 | 0 |
| mel | 0.03 | 0.04 | 0.18 | 0 | 0.45 | 0.29 | 0.01 |
| nv | 0 | 0.02 | 0.1 | 0.01 | 0.11 | 0.75 | 0.01 |
| vasc | 0 | 0 | 0.02 | 0.07 | 0.07 | 0.12 | 0.72 |

Figure 9: Initial Specific Model Classification report and Normalized Confusion Matrix

After fitting this initial model, it is evaluated on the validation data, for which the model an achieved an AUC of 0.91 and an $f_1$ *Score* of 0.66. It is then used to make predictions of the classes of the testing images. To evaluate the model on its performance on the test data set, the metrics in Section 4 are used, the results of which can be seen in Figure 6 and Figure 7.

### 7.1.2 Results

Looking at the top plot of Figure 6, the *ROC curves* for most classes display promising results, given that the *AUC* of all classes are above 0.75, reaching as high as 0.96 for class 'vasc', close to an almost perfect classification for this skin lesion. Moving on to the confusion matrix displayed in Figure 8, it is clear that the model classifies certain classes of lesions better than others, as seen from the fact that 75% of the 'nv' images are classified correctly compared to 'akiec' or 'df', which have a *Recall* of 0.24 and 0.27 respectively. This is potentially due to the much larger amount of 'nv' images the model is trained on.

Finally, analysing the classification report in Figure 8, the weighted average $f_1$ *score* of all the classes is 0.67 and the geometric mean is 0.72, which can be considered relatively adequate as an initial model. One must however note that there is a clear distinction between the $f_1$ *score* and *G-Mean* of classes, with class 'nv' and 'vasc' greatly surpassing the rest. These relatively decent average scores must thus be taken with a pinch of salt. Same holds for metrics such as *Precision.* To try and improve the overall performance of the model in classifying well all forms of skin lesions, we tune certain hyperparameters of the CNN using hyperband tuning.

## 7.2 Tuned Model

### 7.2.1 Model Building/Tuning

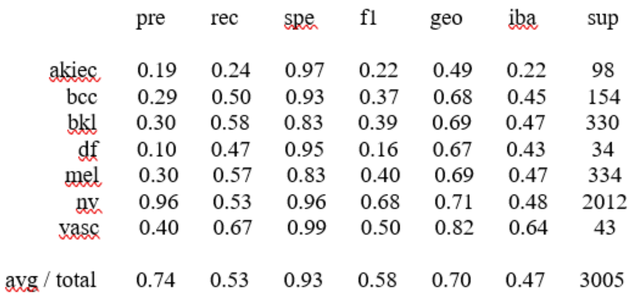
When developing the tuned specific model, hyperband tuning was once again used with the same hyperparameters and ranges as in the general model, given the promising look of the achieved results. The only hyperparameter not tuned in the specific model is that of *binary accuracy*, a natural exclusion based on the multi-class nature of the data used in developing the Specific Model. The rest structure of the model was left untouched and so had the configuration of Figure 2.

Using the maximisation of the Validation *AUC* as an objective, the tuner identified the optimal hyper parameters to be Learning rate: 0.0001, Dropout Rate: 0.1 and Number of Units: 32.

Using these parameters the model was then trained for 80 epochs, and evaluated on the testing data. Examples of predictions can be seen in Figure 12 found in the Appendix.

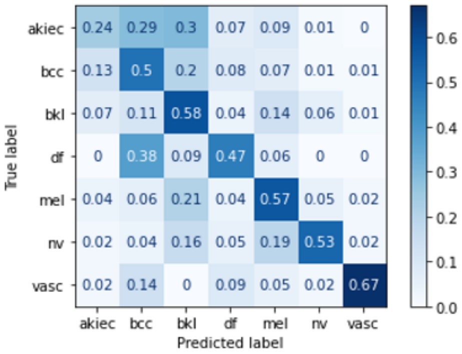
### 7.2.2 Results

Below one can see the classification report and confusion matrix, developed when making predictions on the test images.



| | pre | rec | spe | f1 | geo | iba | sup |
|---|---|---|---|---|---|---|---|
| akiec | 0.19 | 0.24 | 0.97 | 0.22 | 0.49 | 0.22 | 98 |
| bcc | 0.29 | 0.50 | 0.93 | 0.37 | 0.68 | 0.45 | 154 |
| bkl | 0.30 | 0.58 | 0.83 | 0.39 | 0.69 | 0.47 | 330 |
| df | 0.10 | 0.47 | 0.95 | 0.16 | 0.67 | 0.43 | 34 |
| mel | 0.30 | 0.57 | 0.83 | 0.40 | 0.69 | 0.47 | 334 |
| nv | 0.96 | 0.53 | 0.96 | 0.68 | 0.71 | 0.48 | 2012 |
| vasc | 0.40 | 0.67 | 0.99 | 0.50 | 0.82 | 0.64 | 43 |
| avg / total | 0.74 | 0.53 | 0.93 | 0.58 | 0.70 | 0.47 | 3005 |

Figure 10: Tuned Specific Model Classification report



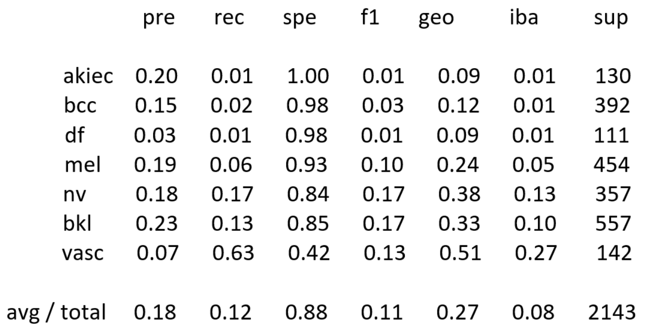
| | akiec | bcc | bkl | df | mel | nv | vasc |
|---|---|---|---|---|---|---|---|
| akiec | 0.24 | 0.29 | 0.3 | 0.07 | 0.09 | 0.01 | 0 |
| bcc | 0.13 | 0.5 | 0.2 | 0.08 | 0.07 | 0.01 | 0.01 |
| bkl | 0.07 | 0.11 | 0.58 | 0.04 | 0.14 | 0.06 | 0.01 |
| df | 0 | 0.38 | 0.09 | 0.47 | 0.06 | 0 | 0 |
| mel | 0.04 | 0.06 | 0.21 | 0.04 | 0.57 | 0.05 | 0.02 |
| nv | 0.02 | 0.04 | 0.16 | 0.05 | 0.19 | 0.53 | 0.02 |
| vasc | 0.02 | 0.14 | 0 | 0.09 | 0.05 | 0.02 | 0.67 |

Figure 11: Tuned Specific Model Normalized Confusion matrix

The *ROC curves* of all classes as displayed in the bottom plot of Figure 6 remain relatively unchanged when comparing to the initial model, all having good shape and *AUC's* well above 0.5, again indicating satisfactory results. The confusion matrix in Figure 9 indicates that the recall of many classes, in particular 'bcc', 'df', 'df' and 'mel' have all increased, however at the slight detriment of the majority class 'nv'. The overall performance of the model, seems to have fallen with lower macro average $f_1$ *score* and $G - Mean$ as seen in Figure 8. This however might be deemed to be the result of the model not performing as well predicting the majority class. Although this might seem bad on first sight, it is important to note, that this model has a more balanced performance across all classes, as well as performs better in predicting classes relating to malignant diseases such as melanoma (mel) and basal cell carcinoma (bcc), for which most metrics have increased. Given the importance of reducing the number of false negatives for these two classes, the fact that many metrics and in particular *Recall* has increased to 0.5 and 0.57 respectively, although still not very good values, indicates that the tuned model might be considered a better choice.

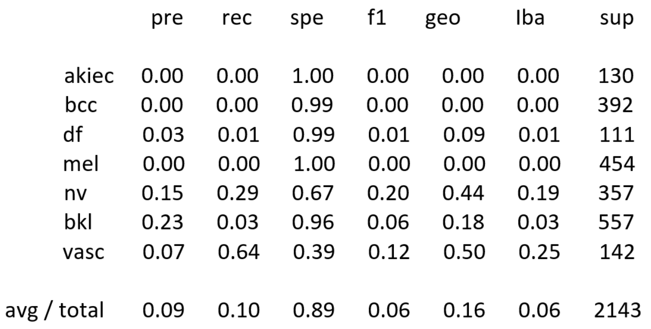
## 8. Application of Specific models to unseen data set

Now that the models have been constructed, let us apply them now to the new data set containing the labels for the different types of cancers described in Section 2. Our expectations should not be very high as the performance of the models on the test data were not good enough. Furthermore, the models are applied now to new data, which which may be different to the images used for the training and so this can result in a harder task for the model to adequately predict the true label of the images.



| | pre | rec | spe | f1 | geo | iba | sup |
|---|---|---|---|---|---|---|---|
| akiec | 0.20 | 0.01 | 1.00 | 0.01 | 0.09 | 0.01 | 130 |
| bcc | 0.15 | 0.02 | 0.98 | 0.03 | 0.12 | 0.01 | 392 |
| df | 0.03 | 0.01 | 0.98 | 0.01 | 0.09 | 0.01 | 111 |
| mel | 0.19 | 0.06 | 0.93 | 0.10 | 0.24 | 0.05 | 454 |
| nv | 0.18 | 0.17 | 0.84 | 0.17 | 0.38 | 0.13 | 357 |
| bkl | 0.23 | 0.13 | 0.85 | 0.17 | 0.33 | 0.10 | 557 |
| vasc | 0.07 | 0.63 | 0.42 | 0.13 | 0.51 | 0.27 | 142 |
| avg / total | 0.18 | 0.12 | 0.88 | 0.11 | 0.27 | 0.08 | 2143 |

Figure 12: Initial Specific Model on unseen data Classification Report



| | pre | rec | spe | f1 | geo | Iba | sup |
|---|---|---|---|---|---|---|---|
| akiec | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 130 |
| bcc | 0.00 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 392 |
| df | 0.03 | 0.01 | 0.99 | 0.01 | 0.09 | 0.01 | 111 |
| mel | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 454 |
| nv | 0.15 | 0.29 | 0.67 | 0.20 | 0.44 | 0.19 | 357 |
| bkl | 0.23 | 0.03 | 0.96 | 0.06 | 0.18 | 0.03 | 557 |
| vasc | 0.07 | 0.64 | 0.39 | 0.12 | 0.50 | 0.25 | 142 |
| avg / total | 0.09 | 0.10 | 0.89 | 0.06 | 0.16 | 0.06 | 2143 |

Figure 13: Tuned Specific Model on unseen data Classification Report

### 8.1 Initial Model

The results of the initial model after trying to predict new data are displayed in Figure 3.6. It can be seen there is a clear drop in most of the metrics when testing on this new data set, having an average of *Precision*, *Recall*, $F_1$*-Score* and *G-Mean* a decrease of 74%, 82%, 84% and 63%, respectively. The only average metric that increases is the *Specificity* by 7%. The best *Recall* in the initial model was obtained by the class 'vasc', which is quite surprising as it is one of the classes with the lowest number number of training observations. However, this is due to the *Specificity* being very low and implying that the *FPR* is 0.58, which results in having most of the observations wrongly classified as 'vasc'(this is also indicated by the low value of the *Precision*). The model is not trained enough for certain classes, however for 'nv', even though it is the class that had the highest number of images to train the model it does not have the highest *Sensitivity* nor the highest *Precision*.

### 8.2 Tuned Model

The application of the tuned model to unseen data results in even worse in terms of evaluation metrics than the initial model. It has a *Recall* and a *Precision* of 0 for the classes 'akiec' and 'bcc', which implies the model does not predict any image to belong to one of these classes. The average of the metrics of the tuned model for test data drop approximately by 88%, 81%, 4%, 90%, 77%, 87% for the *Precision*, *Recall*, *Specificity*, $F_1$*-Score* and *G-Mean* after applying it to new data, respectively.

### 8.3 Problems and tricks to mitigate the performance drop

Similarly to what we saw for the General model predicting the labels for images of new data is a hard problem as they might be different and also may have different features or a different feature distribution. Eliminating possible outliers and through feature selection can mitigate this problem and avoid overfitting.

Furthermore, it can be noted that this data set is also imbalanced, however less imbalanced than the data we used for training the model, as the majority and minority class for this new data constitute 26% and 5% of the total number of images, respectively. While in the data used for training the majority and minority class constitute the 70% and 1% of the total number of images. For this reason, it may be a good idea to use this data set for training as it may give a more stable model and better results when predicting for new data with similar type of images. Another way to deal with this imbalancedness is through oversampling, which re-samples less frequent images to adjust their number to predominant classes, or through undersampling the majority class as it constitutes 70% of the total data when training the model.

Likewise to the problems for the General model, the Specific model lacks of enough training for certain classes. Through the use of techniques such as Data Augmentation and Dropout rate might mitigate the performance drop. It was however used when training the models, and it did not result in good performance metrics. Once more increasing this Droptout rate in the training phase may lead to better performance when applying the model to a new data set, as overfitting needs to be avoided due to its negative impact when the model predicts the labels for unseen images.

The complexity of the model is again an issue can also lead to overfitting and to having a bad performance when applied to new data. This can be seen in the application of the tuned model which results in worse performance than the initial model. The model should remain broad enough in order to classify correctly unseen images from new data. In our case, the complexity of CNN is more suitable for the data used in the training rather than for the new data (given the similar features), which results in bad performance for both models, being worse the tuned one. This simpler model can be accomplished by a different choice of the hyperparameters, by reducing the number of epochs, the batch size, the depth of the neural network and the number of nodes per layer; and increasing the learning rate. Also a different choice of the optimizer might improve our results when applying the model on new data.

## 9. Conclusion

To summarize, in this paper there were two models under consideration for skin cancer image classification, a more generic one which classified the images as 'malignant' or 'benign' and a specific model which classified the images by cancer type. Being the first model a better classifier than the second one, however not being able to obtain the optimal results expected, even after intensive training and tuning of the parameters. Furthermore, these models were then used to classify the images for new data sets, which resulted in even more disappointing results due to several factors like the imbalancedness of the training data, the overfitting of the training model and the difference between the features of the training and new data set.

The use of simpler models by changing the values of the hyperparameters, the use of a suitable data set in the training phase and applying pre-processing data techniques (Data Augmentation, oversampling, undersampling and ensemble methods) may be able to mitigate these problems and lead to better performances when the model is tested on a new data set. This is left for further research, as well as the challenge of developing better methods for classifying these skin cancer images. The problem proposed in this study it has a high technical complexity and although the chosen technology has helped in getting better results than by using other algorithms, the results are far from ideal and should not be taken lightly as it is the patient's health that is put at risk.

## References

Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee, 2017.

David R Bickers, Henry W Lim, David Margolis, Martin A Weinstock, Clifford Goodman, Eric Faulkner, Ciara Gould, Eric Gemmen, and Tim Dall. The burden of skin diseases: 2004: A joint project of the american academy of dermatology association and the society for investigative dermatology. *Journal of the American Academy of Dermatology*, 55(3):490–500, 2006.

Jason Brownlee. What is the difference between a batch and an epoch in a neural network? *Machine Learning Mastery*, 20, 2018.

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 821–826. IEEE, 2017.

Eqbal Dohan Chalap and Ghaidaa Raheem Lateef Al-Awsi. A general overview of the genetic effects of extracellular polymers for enterococcus faecium in cancer cells. 2019.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

Naomi Chuchu, Jacqueline Dinnes, Yemisi Takwoingi, Rubeta N Matin, Susan E Bayliss, Clare Davenport, Jacqueline F Moreau, Oliver Bassett, Kathie Godfrey, Colette O'Sullivan, et al. Teledermatology for diagnosing skin cancer in adults. *Cochrane Database of Systematic Reviews*, (12), 2018.

Emma Craythorne and Firas Al-Niami. Skin cancer. *Medicine*, 45(7):431–434, 2017.

Antonia Creswell, Kai Arulkumaran, and Anil A Bharath. On denoising autoencoders trained to minimise binary cross-entropy. *arXiv preprint arXiv:1708.08487*, 2017.

Xinyang Deng, Qi Liu, Yong Deng, and Sankaran Mahadevan. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, 340:250–261, 2016.

Thomas L Diepgen and V Mahler. The epidemiology of skin cancer. *British Journal of Dermatology*, 146:1–6, 2002.

Pratik Dubal, Sankirtan Bhatt, Chaitanya Joglekar, and Sonali Patil. Skin cancer detection and classification. In *2017 6th international conference on electrical engineering and informatics (ICEEI)*, pages 1–6. IEEE, 2017.

Richard Elderman. *Exploring Improvements for Gradient Descent Optimization Algorithms in Deep Learning.* PhD thesis, 2019.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

X Fang, H Luo, and J Tang. Structural damage detection using neural network with learning rate improvement. *Computers & structures*, 83(25-26):2150–2161, 2005.

Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. Artificial intelligence-based image classification for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine*, page 104065, 2020.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview, 2020.

Tianmei Guo, Jiwen Dong, Henjian Li, and Yunxing Gao. Simple convolutional neural network on image classification. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 721–724. IEEE, 2017.

Gery P Guy Jr, Steven R Machlin, Donatus U Ekwueme, and K Robin Yabroff. Prevalence and costs of skin cancer treatment in the us, 2002- 2006 and 2007- 2011. *American journal of preventive medicine*, 48(2):183–187, 2015.

Zhe Hui Hoo, Jane Candlish, and Dawn Teare. What is an roc curve?, 2017.

Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

Saurabh Karsoliya. Approximating number of hidden layer neurons in multiple hidden layer bpnn architecture. *International Journal of Engineering Trends and Technology*, 3(6):714–717, 2012.

LR Klooster. Approximating differential equations using neural odes. B.S. thesis, University of Twente, 2021.

Howard K Koh. Cutaneous melanoma. *New England Journal of Medicine*, 325(3):171–182, 1991.

Matthias Kohl. Performance measures in binary classification. *International Journal of Statistics in Medical Research*, 1(1): 79–81, 2012.

Ho Tak Lau and Adel Al-Jumaily. Automatically early detection of skin cancer: Study based on nueral netwok classification. In *2009 International Conference of Soft Computing and Pattern Recognition*, pages 375–380. IEEE, 2009.

De Salvo Rostamizadeh Talwalkar Li, Jamieson. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18:1–52, 2018.

Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

C Manliguez. Generalized confusion matrix for multiple classes. *URL https://www. researchgate. net/publication/310799885_Generalized_Conf usion_Matrix_for_Multiple_Classes, DOI*, 10, 2016.

Roman C Maron, Michael Weichenthal, Jochen S Utikal, Achim Hekler, Carola Berking, Axel Hauschild, Alexander H Enk, Sebastian Haferkamp, Joachim Klode, Dirk Schadendorf, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer*, 119:57–65, 2019.

Caren Marzban. The roc curve and the area under it as performance measures. *Weather and Forecasting*, 19(6):1106–1114, 2004.

Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, and Jascha Sohl-Dickstein. Learned optimizers that outperform sgd on wall-clock and test loss. *arXiv preprint arXiv:1810.10180*, 2018.

Patrick A Oberholzer, Damien Kee, Piotr Dziunycz, Antje Sucker, Nyam Kamsukom, Robert Jones, Christine Roden, Clinton J Chalk, Kristin Ardlie, Emanuele Palescandolo, et al. Ras mutations are associated with the development of cutaneous squamous cell tumors in patients treated with raf inhibitors. *Journal of clinical oncology*, 30(3):316, 2012.

Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45, 2008.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

Sebastian Raschka. An overview of general performance metrics of binary classifier systems. *arXiv preprint arXiv:1410.5330*, 2014.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

Nitish Srivastava. Improving neural networks with dropout. *University of Toronto*, 182(566):7, 2013.

Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2020.

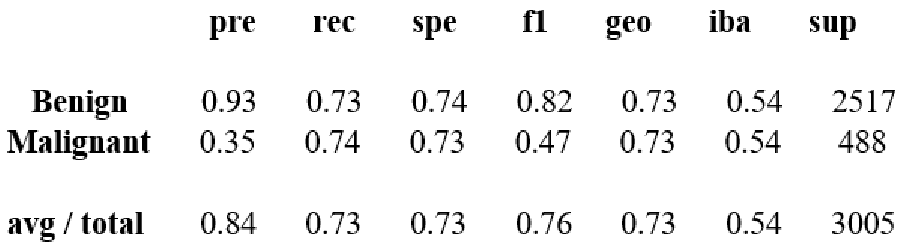
Megan H Trager, Dawn Queen, Faramarz H Samie, Richard D Carvajal, David R Bickers, and Larisa J Geskin. Advances in prevention and surveillance of cutaneous malignancies. *The American journal of medicine*, 133(4):417–423, 2020.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

Nathan E West and Tim O'Shea. Deep architectures for modulation recognition. In *2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pages 1–6. IEEE, 2017.

Min Zhu, Jing Xia, Xiaoqing Jin, Molei Yan, Guolong Cai, Jing Yan, and Gangmin Ning. Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, 6:4641–4652, 2018.

**Appendix A.**



| | pre | rec | spe | f1 | geo | iba | sup |
|---|---|---|---|---|---|---|---|
| **Benign** | 0.93 | 0.73 | 0.74 | 0.82 | 0.73 | 0.54 | 2517 |
| **Malignant** | 0.35 | 0.74 | 0.73 | 0.47 | 0.73 | 0.54 | 488 |
| **avg / total** | 0.84 | 0.73 | 0.73 | 0.76 | 0.73 | 0.54 | 3005 |

Figure 14: Classification Report of the General Model (Section 5.2)

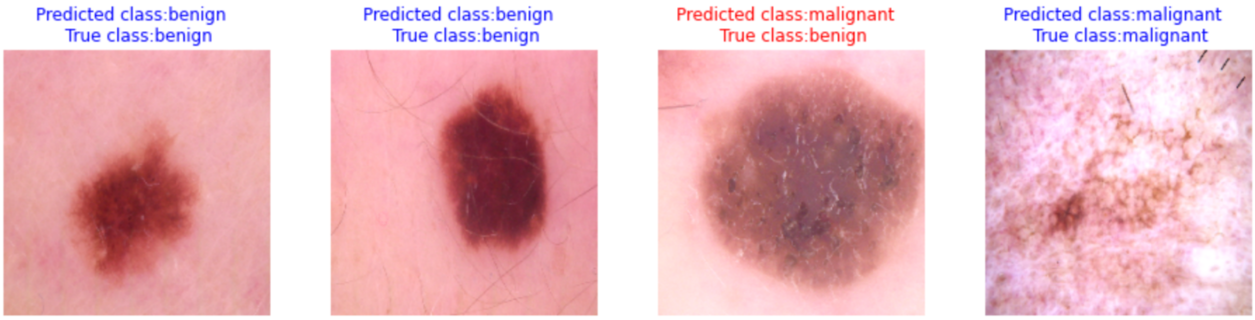


Figure 15: Visualisation of General Model (Section 5.2) on the testing data

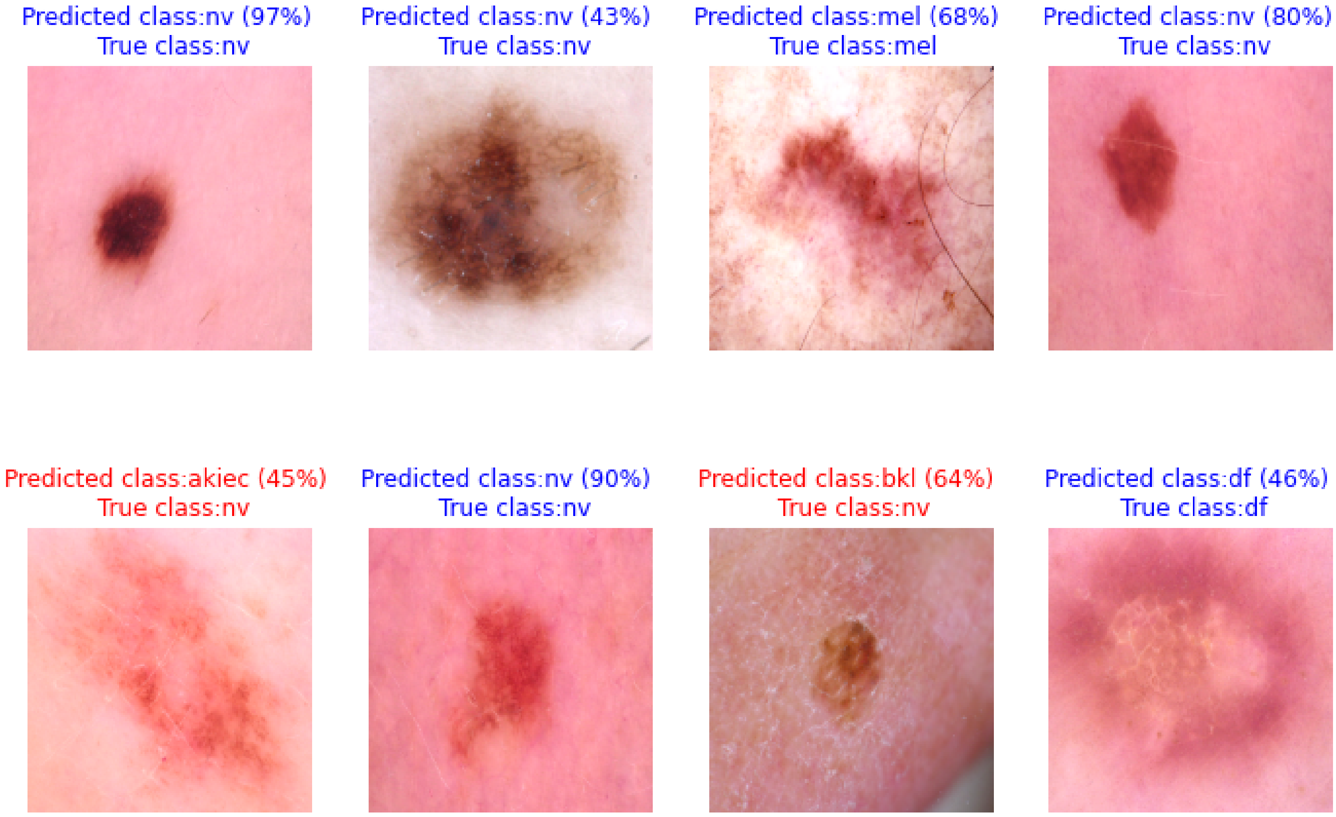


Figure 16: Visualizing tuned specific model predictions on test images of base data set